Journal of Biomedical Informatics 43 (2010) 613–622

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Learning Bayesian networks from survival data using weighting censored instances

Ivan Štajduhar [a,*], Bojana Dalbelo-Bašić [b]

[a] Department of Computing, Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia
[b] Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 29 September 2009
Available online 21 March 2010

Keywords:
Bayesian network
Prognostic model
Survival analysis
Weighting censored instances
Medical decision support

ABSTRACT

Different survival data pre-processing procedures and adaptations of existing machine-learning techniques have been successfully applied to numerous fields in clinical medicine. Zupan et al. (2000) proposed handling censored survival data by assigning distributions of outcomes to shortly observed censored instances. In this paper, we applied their learning technique to two well-known procedures for learning Bayesian networks: a search-and-score hill-climbing algorithm and a constraint-based conditional independence algorithm. The method was thoroughly tested in a simulation study and on the publicly available clinical dataset GBSG2. We compared it to learning Bayesian networks by treating censored instances as event-free and to Cox regression. The results on model performance suggest that the weighting approach performs best when dealing with intermediate censoring. There is no significant difference between the model structures learnt using either the weighting approach or by treating censored instances as event-free, regardless of censoring.

1. Introduction

Standard supervised machine-learning (ML) techniques allow us to learn predictive classification models from data. Once learnt, these models are able to predict whether or not the event of interest (outcome, class) will occur for a new, as yet unclassified instance based on its features. Some of these models express their prediction as a posterior probability distribution.

Survival data, in addition to standard data, is also characterised by survival times. Survival times represent the measured time until an event of interest occurs for that instance. In medicine, an instance is normally described with a patient record and the event of interest usually marks the development of a disease, response to a treatment, relapse, or death of the patient [1]. If the event of interest for a patient was not observed during the follow-up time, that patient's outcome is considered to be censored, and his or her true final outcome is thus unknown.

Censoring in the data is the main reason why standard supervised ML techniques are hard to use for modelling survival. By treating censored data as event-free, one would bias the model towards the event-free outcome. On the other hand, standard statistical techniques for survival analysis [2–4] have no problem at all in dealing with censoring in the data. Moreover, they produce models that are capable of predicting a survival function for a new, as yet unclassified instance. This function expresses the probability of survival, calculated from the instance features, as a function of time.

Estimating survival functions is crucial in domains where the underlying distribution cannot be handled as a classification problem. For example, in palliative patient care, we are certain that death will occur for each and every patient, but would like to estimate when. On the other hand, numerous specialised domains that describe event occurrence until a certain time can be handled as classification problems. For example, if a disease is curable but can also be lethal for a patient if incorrect therapy is applied, determining the final outcome for that patient given his or her known features is of primary interest. A patient that survives the disease will eventually die of other causes, but this should not concern us as it does not imply the need for survival functions. In all cases in which the time to the occurrence of an event is not of interest and observation time equals survival time, virtually any supervised ML technique can learn prognostic models from data [5].

Bayesian networks (BNs) [6] are excellent tools for knowledge representation. They are capable of expressing causal influences probabilistically, which corresponds to human reasoning about causality and uncertainty. They have been enjoying increasing use as decision support systems in various fields of biomedicine and health care [7].

In our previous work [8], we analysed the impact of censoring on learning BNs from survival data by treating censored instances as event-free. We have shown that this data-handling procedure can be efficiently used in the presence of light (up to 20%) censoring. In this paper, we adapted the BN structure-learning algorithms used in [8] by using a censoring weighting scheme

* Corresponding author. Fax: +385 51 651 435.
E-mail address: ivan.stajduhar@riteh.hr (I. Štajduhar).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.03.005

proposed by Zupan et al. [9]. The same procedure was also applied to parameter learning. We compared this procedure with the procedure of learning BNs by treating censored instances as event-free and with Cox regression [2], which is a standard statistical technique for survival analysis.

This paper is organised as follows. Section 2 explains in detail how censored data were preprocessed for their use with the ML algorithms. Section 3 describes the weight-adapted algorithms for learning BNs from data. Related work is described in Section 4. Thorough empirical performance testing in a simulation study is explained next in Section 5, followed by clinical application in Section 6. Section 7 summarises and interprets the results.

2. Handling censored survival data

In this paper, we adopted a procedure for handling censored survival data that divides the data into three groups, as suggested by Zupan et al. [9]: (1) instances for which the event occurred at any time are labelled positive; (2) instances censored after a certain critical time point T^* are labelled negative (event-free); (3) instances censored before time point T^* are doubled, split into both possible outcomes, and then assigned an estimated probability of outcome based on the Kaplan–Meier method. Labelling is illustrated in Fig. 1. Suppose we have a dataset consisting of six instances, $(\mathbf{x}_A^?, \mathbf{x}_B^?, \mathbf{x}_C^?, \mathbf{x}_D^?, \mathbf{x}_E^+, \mathbf{x}_F^+)$, four of which are censored (superscripted ?) and two of which are positive instances (superscripted +). Labelling procedure transforms this dataset into $(\mathbf{x}_A^+, \mathbf{x}_A^-, \mathbf{x}_B^-, \mathbf{x}_C^+, \mathbf{x}_C^-, \mathbf{x}_D^+, \mathbf{x}_D^-, \mathbf{x}_E^+, \mathbf{x}_F^+)$, increasing dataset size by half. All split instances are then assigned weights, based on their estimated outcome probabilities. This is described next.

The Kaplan–Meier product limit estimate [10] of inherent survival function $S(t)$ is given by:

$$\widehat{S}(t) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i}\right) = \widehat{S}(t-1)\left(1 - \frac{d_t}{n_t}\right), \tag{1}$$

where d_i is the number of events that occurred at time t_i (when one or more events occurred) and n_i is the number of patients still observed at time t_i. This method assumes that the censoring times are independent of the survival times. It is not suitable in cases where a record is censored due to reasons related to the causes of event occurrence [1].

Each doubled instance (that was censored before T^*) is assigned weights according to its observation time T. Its negative spawn is assigned weight $w^- = \widehat{S}(T^*)/\widehat{S}(T)$, whereas its positive spawn is assigned weight $w^+ = 1 - w^-$. Given two censored instances, the one observed for a longer time period should have a higher probability of survival until T^* than the one observed for a shorter time period. Time point T^* is usually determined using expert knowledge, depending on the underlying problem. For example, a 5-year interval is considered to be sufficient for follow-up of oncology patients [11–13]. If the underlying survival distribution is exponential in nature, as T approaches T^*, w^- approaches 1, and w^+ approaches 0. As the derivative of the survival function $S\prime(t)$ approaches zero, T^* can be chosen arbitrarily. Since both the simulation study domains and the clinical domain have (approximately) exponentially distributed observation times, we chose to use the largest observation time for T^*. The observation time is discarded from the dataset after applying weights.

Weight assignment is illustrated using the example from Fig. 1. Kaplan–Meier survival estimate for the dataset is presented in Fig. 2. Both positive instances ($\mathbf{x}_E$ and $\mathbf{x}_F$) and the single negative instance ($\mathbf{x}_B$) are assigned weight 1. Since the largest observation time is selected as the critical time point, $T^* = 14$, estimated terminal probability of survival $\widehat{S}(T^*) = 0.55$. Censored instances ($\mathbf{x}_A, \mathbf{x}_C$ and $\mathbf{x}_D$) are split into both outcomes and assigned the following weights: $w(\mathbf{x}_C^-) = 0.55/1 = 0.55, w(\mathbf{x}_C^+) = 1 - w(\mathbf{x}_C^-) = 0.45, w(\mathbf{x}_A^-) = 0.55/0.8 = 0.6875, w(\mathbf{x}_A^+) = 1 - w(\mathbf{x}_A^-) = 0.3125, w(\mathbf{x}_D^-) = 0.55/0.55 = 1, w(\mathbf{x}_D^+) = 1 - w(\mathbf{x}_D^-) = 0$.

This procedure for weighting censored records was used for learning Bayesian networks. Following data pre-processing and weight assignment, the modified dataset can be used for learning both network structure and network parameters. Details regarding the use of weighted data by the BN learning algorithms is described in Section 3.1 and Section 3.2. Cox regression, on the other hand, did not need any intervention in handling censored survival data for learning. Some of the evaluation metrics did, however, require the projection of the estimated survival functions onto single probabilities of survival. For this purpose we chose the probabilities of survival at median observation time.

3. Bayesian networks

Bayesian networks are a powerful formalism for knowledge representation and reasoning under uncertainty. They encode conditional independence relationships among vertices and can be used to represent causal interactions. They are, however, incapable of modelling time-variant covariate interactions. Unlike standard BNs (also referred to as static), dynamic BNs handle temporal relationships among covariates [14]. This can be extremely useful for reasoning about time in tasks such as diagnosis, prognosis, and

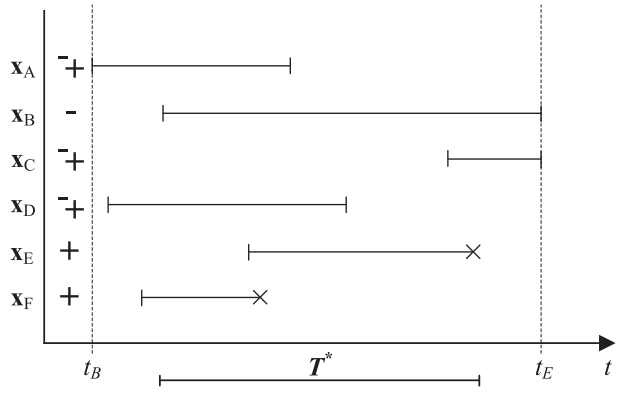


Fig. 1. Labelling instances based on their T/T^* proportion (observation time/critical time). Instances $\mathbf{x}_E$ and $\mathbf{x}_F$ are originally positive (observations ending with an X), other are censored. Labels are assigned in the following way: (1) positive ($\mathbf{x}_E$ and $\mathbf{x}_F$); (2) negative ($\mathbf{x}_B$); (3) split into both possible outcomes ($\mathbf{x}_A, \mathbf{x}_C$ and $\mathbf{x}_D$).

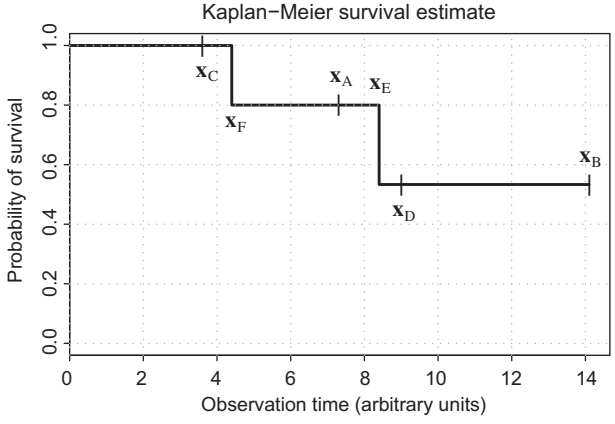


Fig. 2. Kaplan–Meier survival estimate for the sample from Fig. 1. Observation time endpoints of all instances are marked on the survival curve.

(1) $A(G) \leftarrow \emptyset$
(2) estimate $\hat{P}r(G|D)$
(3) $minscore \leftarrow MDL(B, D), lastminscore \leftarrow \infty$
(4) while $minscore < lastminscore$
  (4.1) $lastminscore \leftarrow minscore$
  (4.2*) for every single legal atomic operation on G
    (4.2.1) estimate $\hat{P}r(G'|D)$
    (4.2.2) $score \leftarrow MDL(B', D)$
  (4.3) choose network structure G' with lowest $score$
  (4.4) if $score < minscore$ then
    (4.4.1) $minscore \leftarrow score$
    (4.4.2) $G \leftarrow G'$

*Possible atomic operations ($\forall i \forall j, i \neq j$):
1. arc addition: $A(G') \leftarrow A(G) \cup (V_i \rightarrow V_j)$
2. arc removal: $A(G') \leftarrow A(G) \backslash (V_i \rightarrow V_j)$
3. arc reversal: $A(G') \leftarrow A(G) \backslash (V_i \rightarrow V_j) \cup (V_i \leftarrow V_j)$

Fig. 3. Pseudo code for the hill-climbing algorithm (HC) for learning BN structure.

treatment options [6]. In this paper, we concentrate solely on the application of static BNs.

A Bayesian network B [15] is formally defined as a pair $B = (G, Pr)$, where G is a directed acyclic graph $G = (V(G), A(G))$ with a set of vertices $V(G) = \{V_1, V_2, \ldots, V_N\}$, representing stochastic covariates, and a set of arcs $A(G) \subseteq V(G) \times V(G)$, representing conditional and unconditional stochastic (in)dependencies among the covariates. On the set of covariates V, a joint probability distribution Pr is defined that respects the independencies represented in the graph: $Pr(V_1, \ldots, V_N) = \prod_{i=1}^{N} Pr(V_i|\pi(V_i))$, where $\pi(V_i)$ stands for the covariates corresponding to the parents of the vertex V_i. Learning BNs from data involves two subtasks: (1) learning network structure (determining dependencies, qualitative part) and (2) learning the parameters (determining the strength of these dependencies, quantitative part).

3.1. Structure learning from censored data

Most methods for learning BN structures from data are either score-based or constraint-based. Score-based methods search for the model structure G that best matches the data D by introducing a scoring function that evaluates each model candidate with respect to D [16–19]. Commonly used scoring functions include belief scoring functions [17] and minimum description length-based scoring functions (MDL) [20]. Constraint-based methods, on the other hand, use conditional independence statements (constraints) that are determined by statistical tests on the data [21–23].

When working with discrete data, both the scoring function (score-based methods) and conditional independence tests (constraint-based methods) are calculated via frequency distributions over conditional subspaces [24]. Censored survival data, preprocessed by the method described in Section 2, are easily handled by both BN structure-learning methods by taking into account the weight distribution (w^-, w^+) for the class bin counts. Both conditional and unconditional class probability distributions are then calculated as the average of weights $\sum w^{\{-,+\}} / \sum w$, instead of instance counts.

For testing the described procedure, we used a free data mining software Weka [25]. Although many of Weka's implemented ML algorithms normally handle weighted data, this is not the case with its implementations of BN learning algorithms, so we had to make some adjustments to the software. Next, we describe two standard and well-known algorithms for learning BNs from data, one representing the score-based methods, and the other representing the constraint-based methods.

3.1.1. Hill-climbing algorithm

A scoring function for a Bayesian network $B = (G, Pr)$ defined with structure G and parameters Pr is determined using the network's data likelihood $\mathscr{L}(D|B) = \mathscr{L}(D|G, Pr)$. To prevent overfitting, the score is modified by adding a factor to penalise overly complex structures. The weight-adapted MDL scoring function [20] is used as the criterion function to be minimised. It is given by the following equation:

(1) $A(G) \leftarrow$ completely connected graph
(2) for each connected vertex pair (V_i, V_j)
  (2.1) for each subset $\boldsymbol{Z}$ of neighbours adjacent to both V_i and V_j
    (2.1.1) if $(V_i \perp V_j | \boldsymbol{Z})$ then $A(G) \leftarrow A(G) \backslash (V_i - V_j)$
(3) for each non-adjacent vertex pair (V_i, V_j)
  (3.1) if $V_i - V_k - V_j$ then direct arcs $V_i \rightarrow V_k \leftarrow V_j$
(4) direct the remaining arcs using graphical rules defined in [27]

Fig. 4. Pseudo code for the conditional independence algorithm (CI) for learning BN structure.

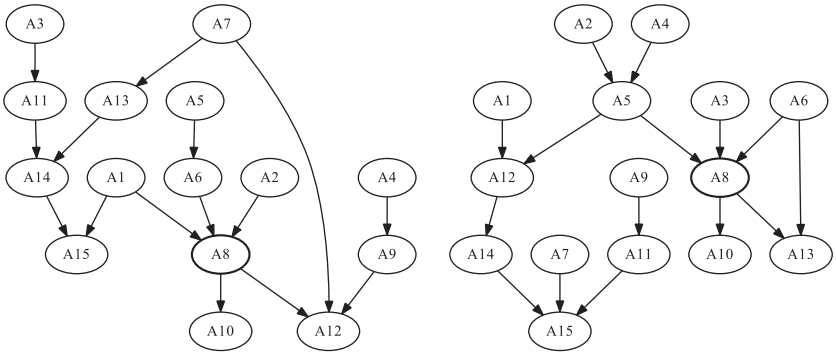


Fig. 5. Two randomly generated BNs from the simulation study.

$$\mathrm{MDL}(B,D) = \frac{d}{2}\log_2 N - \log_2 \mathscr{L}(D|G,\hat{Pr}), \tag{2}$$

where d is the number of free parameters of multinomial local conditional probability distribution tables, N is the number of instances in the sample and $\hat{Pr}$ represents the local conditional probability distribution tables estimated from dataset D (see Section 3.2).

Since the space of all possible structures is at least exponential in the number of covariates, exhaustive search is impractical (mostly impossible). A heuristic search procedure is used instead [26]. Pseudo code for the algorithm is presented in Fig. 3 (steps are referenced in the text). The search begins with an empty graph (step 1). For each attribute pair, an attempt is made to add, remove, or reverse an arc (step 4.2). The network that minimises the score (step 4.3) becomes the current candidate (step 4.4.2), and then the process is iterated. The process stops when no single-arc change can further lower the score (step 4). This does not guarantee global optimum convergence.

Table 1
Observation time hazards used for different levels of censoring.

Censoring	10%	20%	30%	40%	50%	60%	70%	80%
λ_C	0.0003	0.0004	0.0006	0.0012	0.002	0.0033	0.0067	0.01

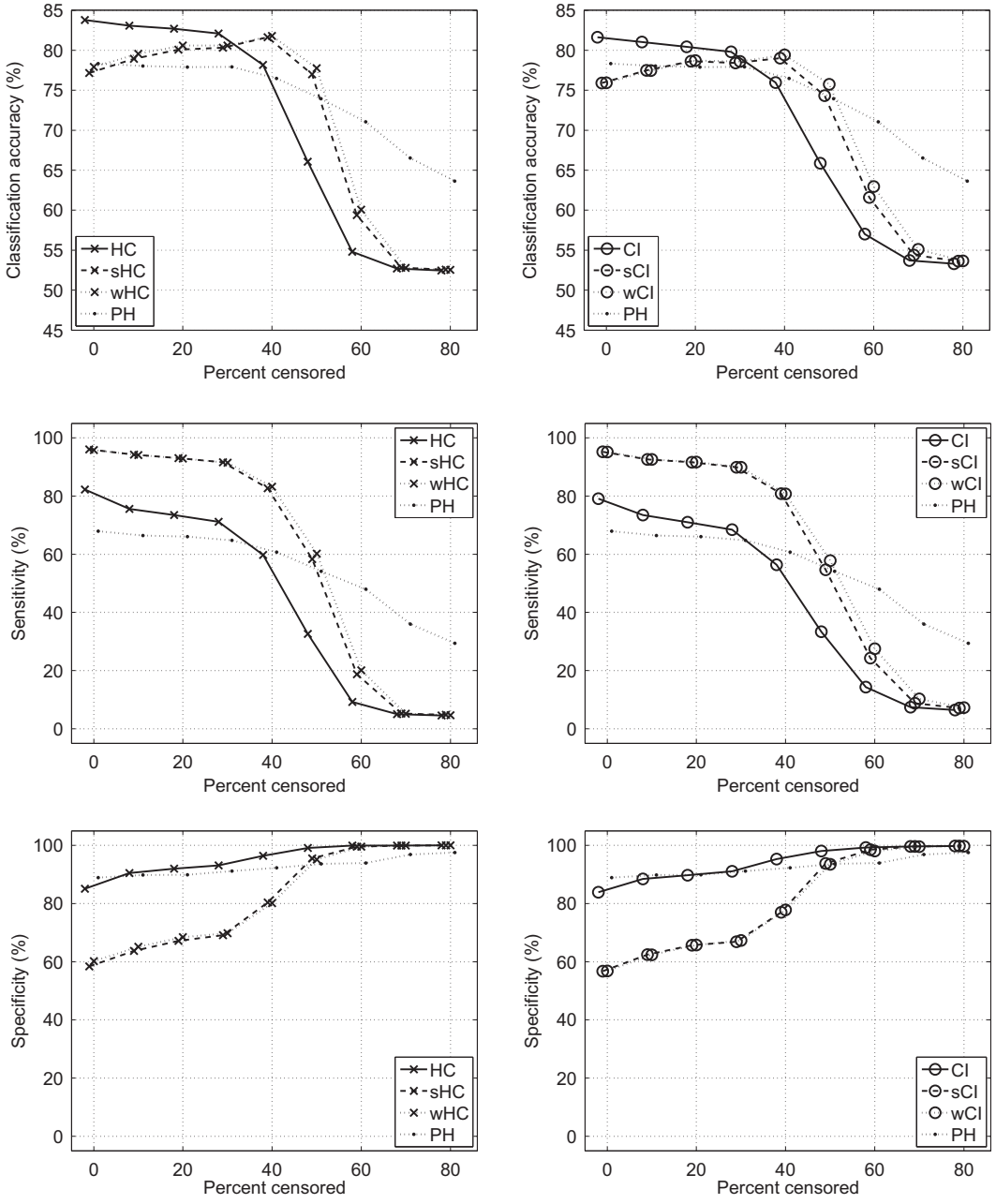


Fig. 6. True classification accuracy, sensitivity and specificity of methods used under different levels of censoring (simulation study).

3.1.2. Conditional independence algorithm

Pseudo code for the conditional independence algorithm is presented in Fig. 4. Starting with a complete undirected graph (step 1), the algorithm tries to find conditional independencies $\langle V_i, V_j | \boldsymbol{Z} \rangle$ in the data. For each pair of covariates (V_i, V_j), it considers subsets of covariates that are neighbours of both V_i and V_j, $\boldsymbol{Z}$, ranging in cardinality from zero to the number of covariates minus two (step 2.1). If an independency is identified, the arc between V_i and V_j is removed from the skeleton (step 2.1.1). To test whether a covariate pair (V_i, V_j) is conditionally independent given a set of covariates $\boldsymbol{Z}$, a network structure with arcs $\forall V_k \in \boldsymbol{Z} : V_k \rightarrow V_j$ is compared with one with arcs $\{V_i \rightarrow V_j\} \cup \forall V_k \in \boldsymbol{Z} : V_k \rightarrow V_j$. The test is done with weight-adapted Bayesian metric [17]. Arc directions are then assigned following a set of graphical rules (steps 3 and 4) [27]. If the data do not have a perfect map [15], then the algorithm will not be able to assign directions for all of the detected arcs.

3.2. Parameter learning from censored data

Once a BN structure is determined, conditional probability tables (CPTs) are estimated directly from the data by calculating frequency distributions over conditional subspaces [24]. As with structure learning, both conditional and unconditional class probability distributions are calculated as the average of weights $\sum w^{\{-,+\}} / \sum w$, instead of instance counts.

4. Related work

Standard ML techniques need to be adjusted to survival data, primarily because of their inherent censoring. The most commonly used approaches for data and method adjustment include (1) treating censored instances as event-free [28], (2) learning separate models from observation time-divided data [29,30], (3) removing instances observed for shorter time periods [11,31] and (4) weighting censored instances [9]. Artificial neural networks [31], decision trees [9] and support vector machines [32] have been successfully applied to different fields of clinical medicine and molecular biology.

On the other hand, little effort has been made to use BNs in traditional survival analysis [7]. We encountered only two papers dealing with BNs for learning from censored survival data. Sierra and Larranaga [33] studied the application of genetic algorithms to learning BNs from data. They handled censoring by using the second approach described above. Marshall et al. [34] used a dynamic BN for handling the time dimension in survival data. They combined a BN with a latent Markov model, thus handling both causal representation and survival events. The approaches described in both papers produced multiple structure or parameter-wise time-dependent BNs.

We, on the other hand, were interested in producing a single BN to model the final outcome. In our previous work [8], we studied the impact of censoring on learning BNs from survival data by treating censored instances as event-free. We have shown that this simple data-handling method can be efficiently used in the presence of light (up to 20%) censoring. The work presented in this paper can be viewed as an extension to [8], suggesting a possible solution for learning BNs in the presence of intermediate (from 40% up to 60%) censoring.

5. Simulation study

Synthetic data were first sampled from BNs with randomly generated structures and CPTs, which should depict relationships typically seen in prognostic factor studies. Different survival and observation times were then generated and assigned to sampled instances to produce censoring.

5.1. Generating synthetic Bayesian networks

Let a BN consist of 15 vertices $(V_1, \ldots, V_o, \ldots, V_{15})$, representing 14 covariates and an outcome V_o. This particular number of covariates was chosen as the result of a learnability-complexity tradeoff. For each $V_i, i < o$, exactly one arc was added; pointing to V_o with probability $p(V_i \rightarrow V_o) = 0.3\bar{3}$ or pointing to any other subsequent vertex with probability $p(V_i \rightarrow V_j) = 0.6\bar{6}, j > i$. For each $V_i, i > o$, exactly one arc was added; pointing from V_o to V_i with probability $p(V_o \rightarrow V_i) = 0.3\bar{3}$ or pointing to any other subsequent vertex with probability $p(V_i \rightarrow V_j) = 0.6\bar{6}, j > i$. Two additional arcs were added between two randomly chosen vertices $V_i \rightarrow V_j, i < j$, to decrease the probability of disconnected graph topologies. This procedure ensures that the resulting network topology is acyclic. In order to make the numbers of possible causes and effects of the outcome variate approximately equal, we chose $o = 8$. Net-

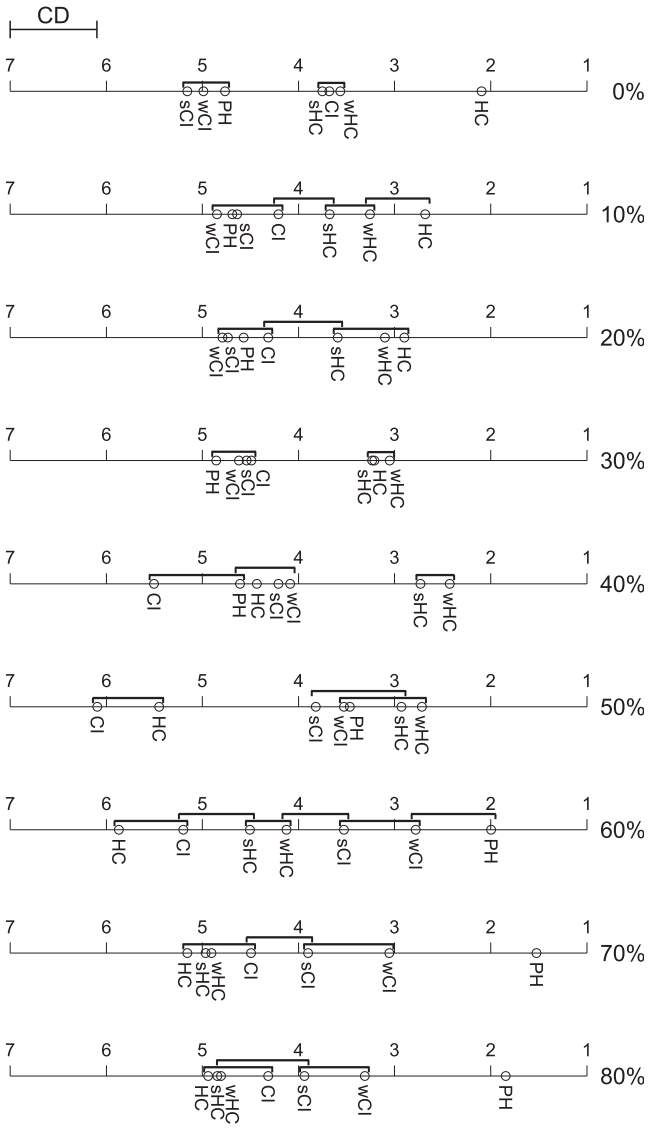


Fig. 7. Average classification accuracy ranks of censored data-handling methods for different levels of censoring (the percentage presented on the right side of each horizontal bar). Groups of methods that are not significantly different ($\alpha = 0.05$) are connected with lines.

works containing any vertices with either four or more parents or four or more siblings were discarded and generated anew. This procedure generates possibly multiply connected yet simple topologies, emphasising direct covariate-outcome interactions. Fig. 5 shows examples of the BN topologies generated.

Next, the interaction of connected vertices was probabilistically expressed via generated CPTs. To simplify the problem, each vertex could assume one of two possible values, either 0 or 1, the latter representing the event-occurred state in the outcome variate. For each orphan V_i, a distribution $(p_i, 1-p_i)$ was generated by sampling p_i from the beta distribution with parameters $\alpha = \beta = 0.2$. For each V_i with m parents, 2^m independent distributions $(p_i, 1-p_i)$ were generated in the same manner, one for each possible combination of parent values. This type of BN will be called a concept BN.

5.2. Sampling and censoring data records

Data were sampled from the generated concept BN, by first ordering the network vertices hierarchically and then sampling a value for each vertex hierarchically, considering the value combination of its parents (if any). To ensure approximately equal quantities of positive and negative records (regarding the outcome),

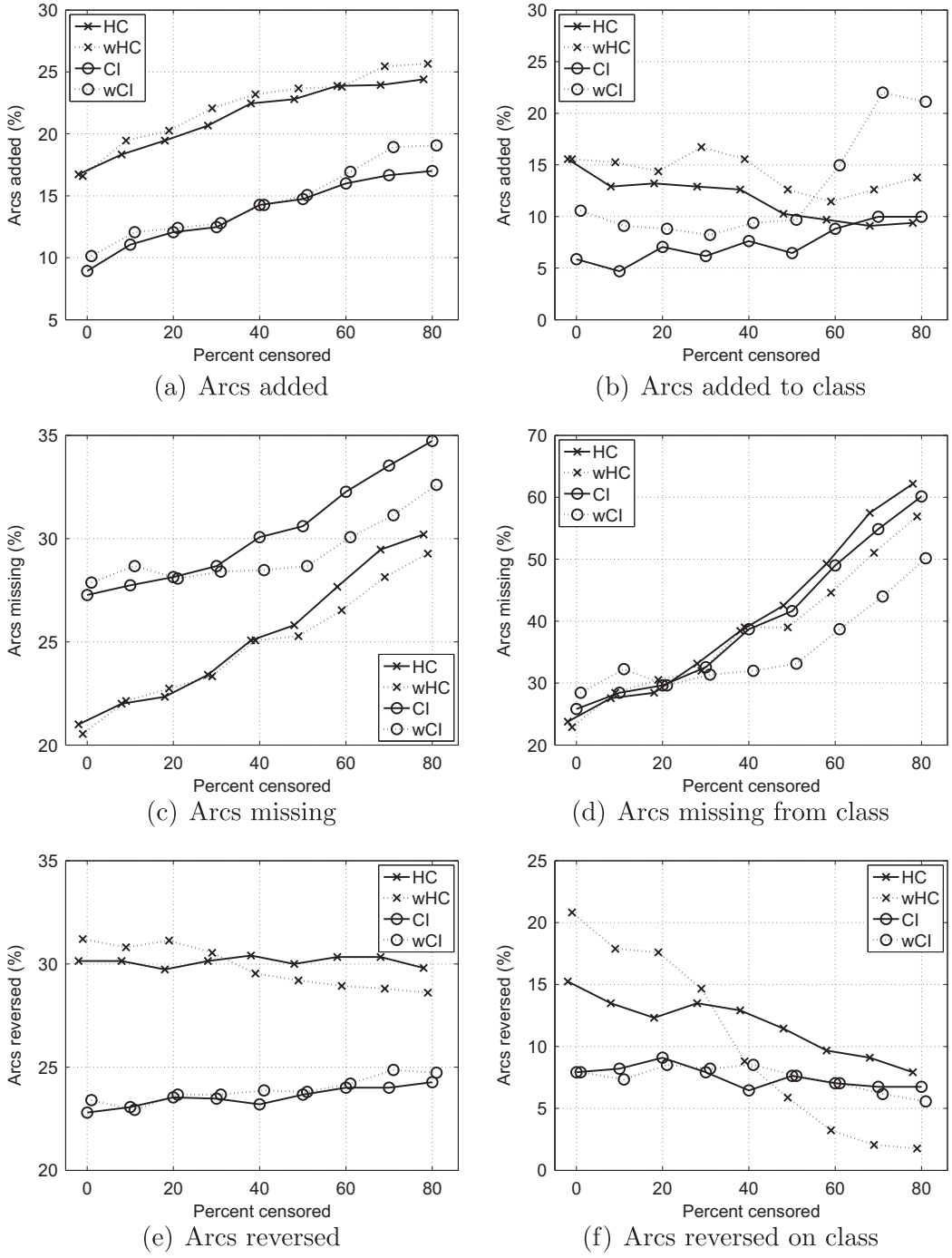


Fig. 8. Structural differences between the concept BNs and learnt BNs, expressed as percentage of total added, missing and reversed arcs (on the left) and percentage of class added, missing and reversed arcs (on the right).

samples with outcome distributions not fitting the constraint $0.45 < p_o < 0.55$ were rejected along with their concept BNs.

Survival and observation times were generated next. Both were simulated using the exponential distribution [35] of the covariates and regression coefficients, estimated from a fitted logistic regression model [36]. The times are expressed by:

$$T_i = -\frac{\ln \upsilon_i}{\lambda \cdot e^{\beta' x_i}}, \tag{3}$$

where υ_i is sampled from a pseudo-random uniform distribution on (0, 1). For survival times we assumed the hazard $\lambda_S = 0.002$. For each simulated survival time, 8 additional pseudo-random, exponentially distributed observation times were generated in the same manner with different hazards λ_C (Table 1), to produce 9 different levels of censoring (from 0% to 80%). Censoring occurs when the generated observation time is shorter than the generated survival time [37]. When that happens, the outcome information reverts from its original state (0 or 1) to a censored state (0).

5.3. Simulation study results

Using the procedure described above, 100 independent BNs were generated. From each model, a separate dataset, containing 1000 instances, was sampled. Each sampled dataset was assigned observation times and then censored at 9 different levels (from 0% to 80%). The results reported here present the average of stratified 10-fold cross-validation over the 100 BNs.

Since we were interested in examining the influence of censoring on both structure and parameter learning of BNs, we compared several censored data-handling setups. In the first setup, we used both algorithms without any intervention, by treating censored instances as event-free (HC, CI). Second, we used the described procedure of handling censored data (Section 2) only on parameter learning, whilst structure learning was performed by handling censored instances as event free (sHC, sCI). Third, we used the procedure for both parameter and structure learning (wHC, wCI). Finally, Cox regression (PH) was also used for reference.

Fig. 6 presents the *true* classification accuracy, sensitivity and specificity of different censored data-handling methods under different levels of censoring. We emphasise the word true, because in this test setup we compared the predicted class values to the original class values, the ones recorded prior to artificial censoring. Handling the censored records as event-free (HC, CI) performs best under light censoring (from 0% to 30%). With intermediate censoring (from 40% do 60%), the weighted method of handling censored data (sHC, wHC, sCI, wCI) outperforms the event-free approach

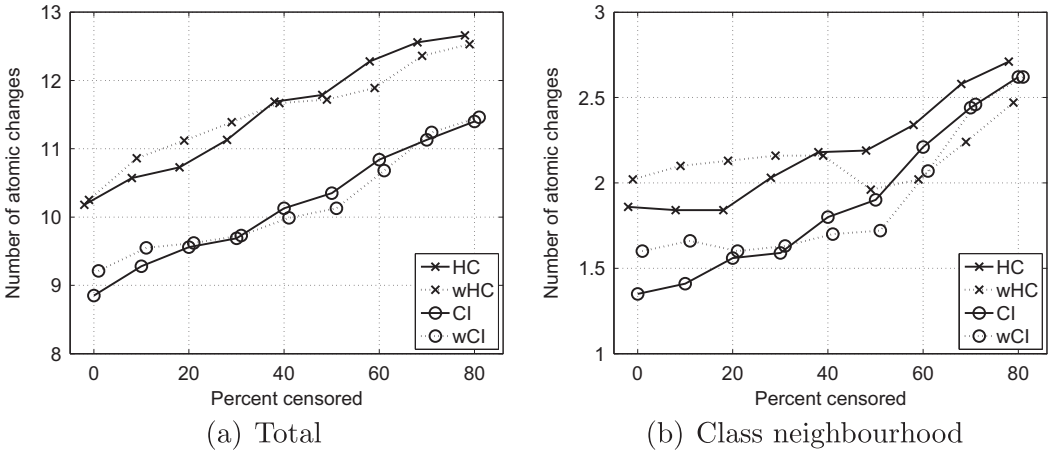


Fig. 9. Number of atomic changes (one arc deletion, insertion or reversal) needed to correct the learnt BN.

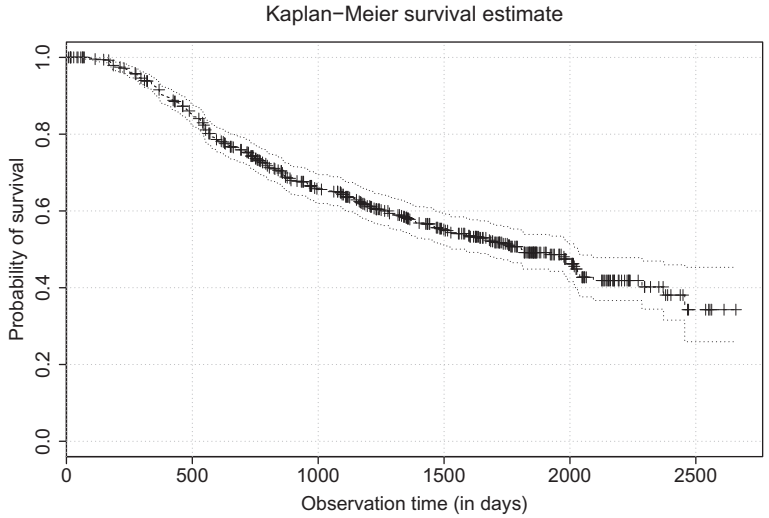


Fig. 10. Kaplan–Meier survival estimate for the GBSG2 dataset. Dashed lines represent the 95% confidence intervals.

(HC, CI). Cox regression (PH) outperforms all approaches under heavy censoring (from 60% to 80%). For both BN learning algorithms, both weighted handling methods increase the sensitivity, but decrease the specificity of the models learnt.

As suggested by Demšar [38], we performed statistical hypothesis tests of performance similarities. For each BN, all methods were ranked according to their classification accuracy, by assigning the best method first place, second best method second place, etc. In the case of a tie, the average rank was assigned (e.g., if the methods ranked 3 and 4 have equal scores, both are assigned the rank 3.5). As was expected, the Iman and Davenport hypothesis test [39] confirmed that the methods are not all equal, regarding classification accuracy. We then performed a post-hoc Nemenyi test [40] for pairwise comparisons of statistical similarities. Two methods are significantly different if the difference of their ranks is larger than the critical difference (CD). For the ranking of 7 methods over 100 independent datasets, under $\alpha = 0.05$, we get $CD = 0.9009$.

Fig. 7 compares the average classification accuracy ranks of different censored data-handling methods for learning Bayesian networks and Cox regression under different levels of censoring. This approximately confirms the results presented in Fig. 6.

Next, we performed tests on the structural differences between the learnt BNs and the concept BNs. Structural difference of two BNs is measured via the number of added (surplus), missing (non-detected) and reversed arcs of (1) the entire network, and (2) the immediate class neighbourhood. These values are then presented as percentages of the total number of arcs in the concept BN (15 arcs) for the first case and percentages of the number of arcs originally connected to the class for the second case (Fig. 8). Two

Table 2
Performance of different censored data-handling methods on the GBSG2 dataset. Values in the table represent the average (with standard deviation) of 10 iterations of stratified 10-fold cross-validation for the following evaluation metrics: weighted classification accuracy (WCA) and concordance index (CInd). The best results are shown in boldface.

Method	WCA (%)	CInd
HC	48.8 (1.2)	0.561 (0.009)
sHC	**65.3** (0.6)	0.573 (0.017)
wHC	65.2 (0.6)	0.570 (0.012)
CI	43.7 (0.2)	0.590 (0.007)
sCI	**65.3** (0.5)	0.593 (0.006)
wCI	65.1 (0.5)	0.546 (0.012)
PH	45.1 (0.3)	**0.651** (0.002)

censored data-handling methods for learning BN structures are compared, one using the weighting scheme (wHC, wCI), and the other treating censored instances as event-free (HC, CI). Surprisingly, there is little difference between the two censored data-handling approaches. Weighting censored records generally produces more surplus arcs and less missing and reversed arcs in the class neighbourhood under intermediate and heavy censoring. Fig. 9 presents the numbers of atomic structure changes needed to correct the learnt BN, so it becomes equal in structure to the concept BN. An atomic change can be the removal, addition, or reversal of one arc. The wHC method performs better than HC under medium and heavy censoring. wCI performs better than CI under medium censoring. This becomes more apparent in the immediate class neighbourhood. Statistical tests inspecting the method performance regarding the number of atomic changes (both in the whole

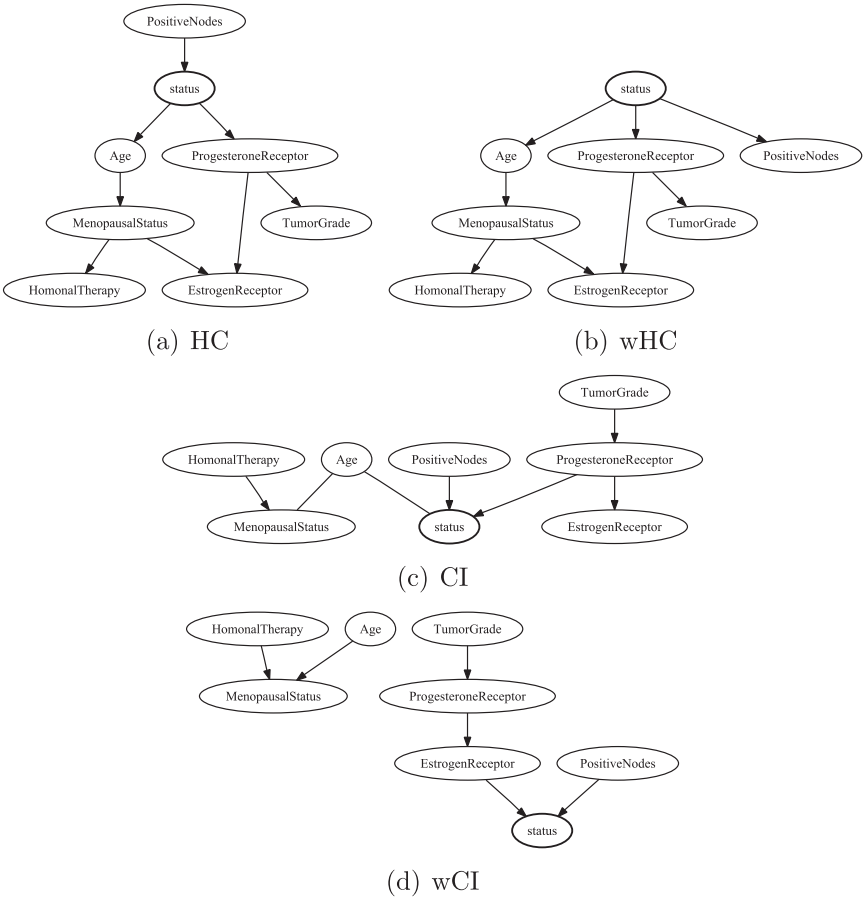
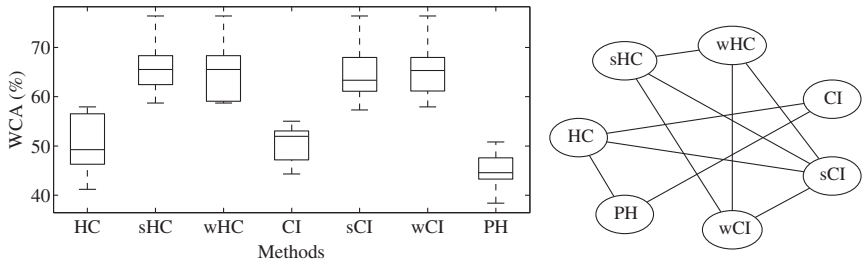
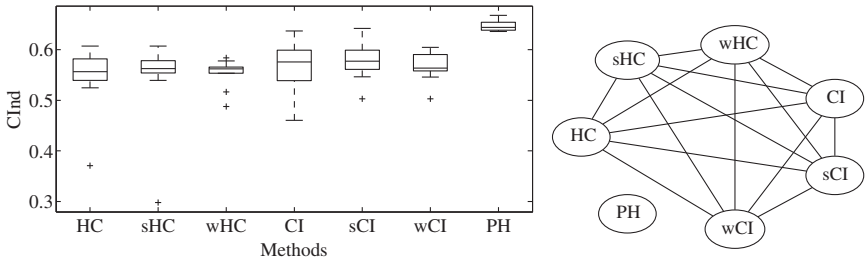


Fig. 11. BN structures learnt with different methods from the GBSG2 dataset.



(a) Weighted classification accuracy



(b) Concordance index

Fig. 12. Box and Whisker plots (on the left) of results on survival evaluation metrics obtained through the 5×2 cross-validation test. Methods not significantly different under $\alpha = 0.05$ are connected with a line (graphs on the right).

network and in the immediate class neighbourhood) did not find any significant difference ($\alpha = 0.05$) between the weighted and event-free approaches to handling censored data. Here we used the same statistical procedure that was used in the classification accuracy tests (Fig. 7).

6. Clinical application

The method performance was next analysed on a clinical dataset. We used data from the study of the treatment of node-positive breast cancer patients, known as the German breast cancer study group (GBSG2) [41]. The phrase node-positive in the name of the disease describes the state, where tumour cells have metastasised to lymph nodes. The dataset includes 686 data records describing the condition of women who had positive regional lymph nodes but no distant metastases. Women included in the study were 65 years of age or younger and were observed for cancer recurrence. The dataset is available as part of the *ipred* package of the *R* environment [42]. It is also available online at http://www.blackwellpublishing.com/rss/Volumes/A162p1.htm (Accessed: 31 December 2009). Kaplan–Meier survival estimate for the GBSG2 dataset is presented in Fig. 10.

Each data record is described with seven prognostic covariates (patient age, menopausal status, tumour size, tumour grade, number of positive nodes, progesterone receptor and oestrogen receptor), a hormonal therapy indicator, observation time, and status. Status describes whether the observation was censored or whether the cancer recurred. Prediction covariates describing patient age, tumour size, number of positive lymph nodes, progesterone receptor levels and oestrogen receptor levels were discretised using the method proposed in [43,44], which is based on determining the best log-rank separation threshold.

BNs learnt using both censored data-handling methods and both BN learning algorithms are presented in Fig. 11.

The performance of the methods was evaluated using two well-known survival evaluation metrics: the weighted classification accuracy (WCA) [31] and concordance index (CInd) [45]. WCA is basically equivalent to the weighting procedure used for learning [9]. Every censored instance is divided into two: a positive one, assigned weight w^+ and a negative one, assigned weight w^- (both are estimated from the Kaplan–Meier survival curve of the test set). Depending on the prediction made, if the outcome is censored, one of the weights increases the count of correct predictions, while the other increases the count of incorrect predictions, by the magnitude of its size. Positive instances have $w^+ = 1$. WCA is then calculated as $100 \cdot \sum w^+ / (\sum w^+ + \sum w^-)$. The concordance index is the probability that, given two randomly selected instances, the instance with the observed worse outcome is predicted to have a worse outcome. It is calculated as a proportion of the consistent instance pairs over the number of usable instance pairs. An instance pair is usable if the event occurred for the instance with a shorter follow-up time. A pair is consistent if the instance with a shorter follow-up time is assigned a higher probability of event occurrence.

The average results (with standard deviations) of 10 iterations of stratified 10-fold cross-validation are presented in Table 2. Weighted handling of censored data for learning BN parameters outperformed other methods, according to WCA. CInd confirms that this method of censored data-handling is better than the others. Cox regression, however, has the highest CInd.

To confirm the significance of the difference in performance between methods, we performed the non-parametric Friedman ANOVA test (which does not make the normal distribution assumption), following the post-hoc Bonferroni adjusted Wilcoxon signed-rank test ($\alpha = 0.05$) [46]. We used the results obtained through 5 iterations of 2-fold cross-validation, to reduce bias [47]. Box and Whisker plots of the tests, accompanied by similarity diagrams, are shown in Fig. 12. The test on WCA suggests that all weighting methods are significantly different ($p < 0.05$) from the non-weighting methods (HC, CI) and Cox regression (PH). The test on CInd suggests that PH is significantly different from HC, wHC, wCI and sHC.

7. Conclusion

Bayesian networks are an excellent tool for knowledge representation. They are capable of acquiring this knowledge from data by using procedures for both structure and parameter learning. BNs combine causal representation of covariate interactions with

their stochastic relationships. Because of their intuitive interpretation, BNs are now widely used as expert models in several different areas of clinical medicine. A number of research papers discussing the problem of using different ML techniques for learning from censored survival data has emerged in the last decade. Only a few of them consider using BNs for modelling survival.

In this paper, we apply the procedure of weighting censored instances [9] for the purpose of learning BNs from censored survival data. We use this weighting procedure both for estimating the parameters (CPTs) from data and for structure learning. Structure learning is performed using an adapted search-and-score hill-climbing algorithm and an adapted constraint-based conditional independence algorithm. The weighting procedure is thoroughly benchmarked against other methods in a simulation study and on a clinical dataset.

Simulation study results on their classification accuracy suggest that the weighting methods should be used with BNs only when dealing with intermediate censoring in the data (from 40% to 60%). If censoring in the data is light (up to 30%), one should use the original algorithms for learning BNs, by treating censored instances as event-free [8]. Heavy censoring renders all the applied methods inoperable, except for Cox regression. Tests on the structural correctness of the learnt BNs suggest, that there is no significant difference between the weighted method and event-free handling of censored records.

Tests performed on the GBSG2 dataset suggest that the weighting methods should be used only for parameter learning. When applying the same method also to structure learning, the results become slightly worse. The difference between the different data-handling approaches for learning BNs is, however, statistically insignificant.

Acknowledgments

This work was generously supported by the Croatian Ministry of Science, Education and Sports, Project No. MZOŠ 069-0362214-1575 and Project No. MZOŠ 036-1300646-1986.

References

[1] Lee ET, Wang JW. Statistical methods for survival data analysis. 3rd ed. Hoboken, NJ, USA: John Wiley & Sons; 2003.
[2] Cox DR. Regression models and life-tables. J R Stat Soc B (Methodological) 1972;34(2):187–220.
[3] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, CA, USA: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
[4] Kooperberg C, Stone CJ, Truong YK. Hazard regression. J Am Stat Assoc 1995;90(429):78–94.
[5] Lucas P, Abu-Hanna A. Prognostic methods in medicine. Artif Intell Med 1999;15(2):105–19.
[6] Neapolitan RE. Learning Bayesian networks. Upper Saddle River, NJ, USA: Prentice Hall; 2003.
[7] Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med 2004;30(3):201–14.
[8] Štajduhar I, Dalbelo-Bašić B, Bogunović N. Impact of censoring on learning Bayesian networks in survival modelling. Artif Intell Med 2009;47(3):199–217.
[9] Zupan B, Demšar J, Kattan MW, Beck R, Bratko I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. Artif Intell Med 2000;20(1):59–75.
[10] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81.
[11] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34(2):113–27.
[12] Lucas P. Expert knowledge and its role in learning Bayesian networks in medicine. Lect Notes Comput Sci 2001;2101:156–66.
[13] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79(4):857–62.
[14] Murphy KP. Dynamic Bayesian networks: representation, inference and learning, Ph.D. thesis. University of California; 2002.
[15] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, CA, USA: Morgan Kaufman; 1988.
[16] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992;9(4):309–47.
[17] Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn 1995;20(3):197–243.
[18] Chickering DM. Optimal structure identification with greedy search. J Mach Learn Res 2002;3:507–54.
[19] Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Mach Learn 2003;50(1):95–125.
[20] Lam W, Bacchus F. Learning Bayesian belief networks: an approach based on the MDL principle. Comput Intell 1994;10(4):269–93.
[21] Pearl J. Causality: models, reasoning, and inference. Cambridge, UK: Cambridge University Press; 2000.
[22] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. Cambridge, MA, USA: MIT Press; 2000.
[23] Cheng J, Greiner R, Kelly J, Bell D, Liu W. Learning Bayesian networks from data: an information-theory based approach. Artif Intell 2002;137(1–2):43–90.
[24] Borgelt C, Kruse R. Graphical models: methods for data analysis and mining. Chichester, United Kingdom: John Wiley & Sons; 2002.
[25] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco, CA, USA: Morgan Kaufman; 2005.
[26] Russell SJ, Norvig P. Artificial intelligence: a modern approach. second ed. Upper Saddle River, NJ, USA: Prentice Hall; 2002.
[27] Verma T, Pearl J. An algorithm for deciding if a set of observed independencies has a causal explanation. In: Dubois D, Wellman MP editors. Proceedings of the 8th annual conference on uncertainty in artificial intelligence. San Fransisco, CA, USA: Morgan Kaufmann; 1992. p. 323–330.
[28] Snow PB, Smith DS, Catalona WJ. Artificial neural networks in the diagnosis and prognosis of prostate cancer: a pilot study. J Urol 1994;152(5):1923–6.
[29] Jerez-Aragonés J, Gómez-Ruiz J, Ramos-Jiménez G, Muñoz-Pérez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med 2003;27(1):45–63.
[30] Lisboa PJG, Wong H, Harris P, Swindell R. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artif Intell Med 2003;28(1):1–25.
[31] Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. In: Gant V, Dybowski R, editors. Clinical applications of artificial neural networks. Cambridge, UK: Cambridge University Press; 2001. p. 237–55.
[32] Evers L, Messow C-M. Sparse Kernel methods for high-dimensional survival data. Bioinformatics 2008;24(14):1632–8.
[33] Sierra B, Larranaga P. Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches. Artif Intell Med 1998;14(1–2):215–30.
[34] Marshall A, McClean S, Shapcott M, Millard P. Learning dynamic Bayesian belief networks using conditional phase-type distributions. Lect Notes Comput Sci 2000:516–23.
[35] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med 2005;24:1713–23.
[36] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York, NY, USA: Springer; 2001.
[37] Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23(5):723–48.
[38] Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.
[39] Iman RL, Davenport JM. Approximations of the critical region of the Friedman statistic. Commun Stat Theory Methods 1980;9(6):571–95.
[40] Nemenyi P. Distribution-free multiple comparisons, Ph.D. thesis. Princeton University; 1963.
[41] Schumacher M, Bastert G, Bojar H, Hubner K, Olschewski M, Sauerbrei W, et al. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group. J Clin Oncol 1994;12(10):2086–93.
[42] R Development Core Team. R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, AT. Available from: http://www.R-project.org; 2008 [accessed 31.12.2009].
[43] Contal C, O'Quigley J. An application of changepoint methods in studying the effect of age on survival in breast cancer. Comput Stat Data Anal 1999;30(3):253–70.
[44] Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. second ed. New York, NY, USA: Springer; 2003.
[45] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. J Am Med Assoc 1982;247(18):2543–6.
[46] Corder GW, Foreman DI. Nonparametric statistics for non-statisticians: a step-by-step approach. Hoboken, NJ, USA: John Wiley & Sons; 2009.
[47] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895–923.